Nouvelle approche classificatoire appliquée au web, Une validation expérimentale : représentation des sciences de l'information et de la communication sur le web

Eric Boutin, Luc Quoniam, Gabriel Gallezot

Nouvelle approche classificatoire appliquée au web

Une validation expérimentale : représentation des sciences de l'information et de la communication sur le web

Eric Boutin[1], Luc Quoniam[2], Gabriel Gallezot[3]

1 Laboratoire I3M, Université du Sud Toulon Var, BP 132 83957 la Garde Cedex , boutin@univ-tln.fr

2 Institut Ingémédia, Université du Sud Toulon Var, BP 132 83957 la Garde Cedex quoniam@univ-tln.fr

3 Laboratoire I3M, Université de Nice, Urfist PACA-C Université de Nice - Sophia Antipolis, gallezot@unice.fr

Résumé : Le Web est un « é-cosystème » [5] documentaire composé de pages web en interaction. Les moteurs de recherche majeurs restituent, suite à la requête d'un internaute, une information organisée sous forme d'une liste de pages disjointes de laquelle il est difficile de dégager une vision d'ensemble. Différentes familles d'outils ont été créés pour répondre à ce problème : les outils de classification en font partie. Après un état de l'art des outils de classification automatique de corpus web, cette communication présente une méthode de classification originale qui fait ensuite l'objet d'une validation expérimentale.

Mots clés : classification automatique, clustering, sciences de l'information et de la communication, moteur de recherche, thésaurus

1. Introduction

Il existe un certain nombre d'outils proposant une classification automatique de corpus web. Beaucoup offrent en ligne une vitrine de leur technologie et se présentent sous la forme de moteurs ou de méta-moteurs de recherche. Ces outils offrent à leurs utilisateurs, suite à une requête, une classification automatique des thèmes principaux touchant à la requête. Certains outils, tels Exalead poussent cette logique plus loin et vont jusqu'à proposer en plus des clusters une sélection par format (RTF, DOC, PDF), par type de document (audio, vidéo, RSS), par langue, par nom de domaine. Ces outils sont destinés à la reformulation de la requête, à la découverte et l'affinage de problèmes flous et se révèlent être une solution pertinente dans une démarche exploratoire. Ils offrent une grille de lecture qui permet d'appréhender plus facilement un corpus web. Ce type d'approche propose aussi une classification atypique qui mélange des concepts appartenant à des registres différents. Quelle représentation du web présentent ces outils ?
Dans cette communication, nous avons établi un état de l'art des outils de classification automatique de corpus web pour en dégager les caractères saillants, les hypothèses implicites et en caractériser les usages. Cette analyse critique nous permet ensuite de faire ressortir les points d'optimisation possibles de ces outils. Nous proposons alors une approche classificatoire originale qui s'affranchit des limites des outils existants. Cette approche fait l'objet *in fine* d'une validation expérimentale : nous nous sommes intéressés à la classification automatique du domaine des « sciences de l'information et de la communication » tel qu'il transparaît sur le web.

2. Etat de l'art

Tapez la requête « sciences de l'information et de la communication » sous le méta moteur Grokker ; faites de même sous vivisimo, iboogie, exalead. Les résultats obtenus sont présentés annexe 1. Soumettez les résultats obtenus à l'expert du domaine (chercheur en sciences de l'information et de la communication) en lui demandant de se positionner par rapport à la classification automatique proposée par chacun de ces outils. Nous nous sommes livrés à cette petite expérience dans un laboratoire en sciences de l'information et de la communication. Nous avons obtenu des réponses variées dépendant de la sensibilité, de l'orientation scientifique de l'expert interrogé. L'expert est amené à confronter sa représentation humaine et individuelle à la représentation automatique renvoyée par l'outil. De façon générale, les personnes interrogées considèrent que l'ensemble des termes des clusters est en phase avec le champs des SIC même si il manque des éléments, si certains termes ne correspondent à rien. Ce type de classification manque de précision.

Il existe de nombreux outils de classification automatique de corpus web. Ces outils ont pour objectif de procéder à des agrégations de pages web. Cette agrégation est obtenue en mobilisant différentes technologies et peut être exploitée par le moteur de différentes manières.

1. De façon générale le processus d'agrégation se réalise sur deux bases algorithmiques principales : l'analyse de contenu, l'analyse relationnelle ou une combinaison des deux. Dans le premier cas [4], deux documents sont considérés comme proches lorsqu'ils ont des contenus textuels voisins. Deux mots clés sont considérés comme proches lorsqu'ils sont présents dans un nombre significatif de pages web. Dans le second cas, [6], [7], [9], deux documents sont agrégés lorsqu'ils sont reliés l'un à l'autre de façon directe ou indirecte par un ou des liens hypertextes.

2. Les catégories ou cluster ainsi constituées sont présentées à l'internaute ou utilisés par le moteur pour ordonner ses résultats de façon transparente sans que l'utilisateur ne s'en rende compte. Plusieurs cas peuvent se produire :

a) Le moteur de recherche peut présenter une double grille de lecture. A coté de la liste de pages web traditionnellement renvoyée par le moteur, celui-ci propose, dans un menu disjoint, une taxonomie créée à la volée [11] . On retrouve ce principe dans divers outils tels grokker, vivisimo.

b) Le moteur de recherche peut utiliser la catégorisation issue de la classification automatique pour représenter les données en remplaçant la traditionnelle liste de réponses par une autre représentation souvent cartographique. On obtient alors une représentation cartographique synthétique. Kartoo propose ce type de navigation.

c) Les techniques de clustering peuvent être aussi prises en compte par les moteurs de recherche sans apparaître dans les résultats visibles par l'internaute. Dans ce cas, la liste traditionnelle proposée par le moteur reste l'unique moyen de navigation proposée à l'internaute mais cette liste, dans sa constitution, intègre la notion de cluster. Certains travaux expérimentaux se sont intéressés à la construction de nouveaux indicateurs de pertinence pour les moteurs de recherche. Ainsi Benyu et al [1] décrivent un processus dans lequel la pertinence d'une page web n'est plus définie par rapport à la requête de l'internaute mais par la contribution de cette page à l'information contenue dans l'intégralité des pages analysées. L'indicateur de pertinence résultant de ce genre d'algorithme renvoie, dans les premiers résultats, des pages non redondantes (issues de clusters différents) permettant de rendre compte du mieux possible de la diversité des réponses possibles. Dans le travail de Benyu et al., ces clusters sont constitués en utilisant l'analyse textuelle. Ainsi l'internaute découvre dans les premières réponses une grande variété de réponses possible non redondantes.

La présente analyse ne porte pas sur l'intégralité des cas de figure qui ont été rapidement esquissés. Nous allons nous focaliser sur les techniques classificatoires qui doublent la liste des réponses du moteur d'une liste de thèmes destinée à aider l'internaute dans son processus de navigation

Pour systématiser l'analyse, nous avons réalisé une étude comparative de ces principaux outils de classification automatique de corpus web. Une liste non exhaustive est fournie annexe 2. Ceux que nous allons présenter ici sont des méta-moteurs. Cette étude nous a permis d'identifier les caractéristiques principales de chacun d'eux. Nous en avons identifié quatre :

1. Ces outils reposent sur des technologies de classification dynamique de documents compatible temps réel. Il n'y a pas de classification préexistante. La classification est effectuée par le moteur suite à une requête de l'internaute. Le méta moteur récupère les résultats du moteur de recherche, les classent à la volée et restitue sa classification. Ceci est une force en ce sens que l'outil donne l'image, la photographie instantanée au moment de la requête mais c'est surtout une nécessité à l'heure de l'immédiateté. Ceci est une faiblesse dans ce que qu'elle entraîne dans le processus de collecte et de traitement de l'information sous jacent.

2. La question de la représentativité du corpus. Lorsqu'on interroge un moteur de recherche, le nombre de réponses obtenues peut être très important. La requête "sciences de l'information et de la communication" donne sur Google 109 000 réponses en Janvier 2006, 140 000 sur Yahoo. Ce nombre élevé n'est pas le fait simplement de cette requête. Le nombre de réponses d'un moteur de recherche suit une loi zipfienne. Un tiers des requêtes a moins de 10000 réponses, un autre tiers entre 10000 et 1000000 réponses et le dernier tiers plus d'un million de réponses. Le méta moteur, de part son fonctionnement, va récupérer une partie seulement de l'information du moteur de recherche. Les moteurs de recherche ne fournissent suite à une requête qu'au maximum 1000 réponses. C'est largement suffisant lorsqu'il s'agit d'en faire une exploitation manuelle mais lorsqu'on dispose d'un outil de traitement automatique de l'information, il pourrait s'avérer judicieux de disposer d'une liste exhaustive. La question qui se pose est donc celle de la représentativité de l'échantillon de 1000 réponses au maximum récupérées par le méta moteur. Dans la réalité, cette limite théorique des 1000 réponses n'est pas atteinte, les méta-moteurs classificateurs se contentant de 250 voire 500 réponses. Lorsqu'on se satisfait, comme dans le cas de la requête

"sciences de l'information et de la communication", qui nous sert de fil rouge, d'un traitement réalisé sur 1000 pages web alors qu'il y en a 109 000 au total, on effectue un échantillon d'un centième des pages web. Cet échantillon ne possède pas des propriétés qui nous permettent de faire des inférences statistiques sérieuses au niveau de la population totale des 109000 pages. En effet la première partie des résultats ne correspond pas forcément pas à un échantillon aléatoire, ni d'ailleurs aux résultats les plus pertinents mais aux résultats émanant de sites web qui investissent les ressources nécessaires pour être bien positionnés dans les moteurs de recherche. Sauf dans des cas très académiques qui échappent peut être encore à la logique marchande, ces représentations de synthèse renvoient un reflet déformé de la réalité. Il y aurait lieu de considérer non plus un échantillon non représentatif de résultats mais l'ensemble des résultats du moteur de recherche qui est lui-même un échantillon des pages traitant du sujet. Seuls les moteurs majeurs peuvent réaliser ce type de traitement mais ils n'ont pour l'instant pas choisis cette voie. Il y a sur le web une asymétrie du contexte entre un petit nombre d'acteurs (Google, Yahoo, Microsoft) qui ont accès à des bases de données géantes et l'information qu'ils distillent aux autres acteurs. Les choix technologiques d'une réponse dynamique couplée à l'asymétrie du contexte informationnel conduisent donc à un questionnement sur la validité des résultats obtenus par une approche classificatoire telle qu'elle est actuellement menée par les outils étudiés.

3. Les biais dans la source d'information. Les technologies utilisées par les méta-moteurs classificateurs reposent toutes sur l'analyse de contenu. C'est l'analyse textuelle qui va permettre d'identifier des mots caractéristiques de la requête. Le point est ici plutôt de savoir sur quelles informations sont construites ces classifications. Elles sont bâties à partir de l'information transmise par le moteur de recherche au méta moteur. L'analyse textuelle n'est donc pas menée sur la page web entière mais de quelques lignes correspondantes au contenu de la balise titre et d'un résumé de la page web proposé par le moteur de recherche. Les techniques utilisées par les moteurs de recherche pour construire ce résumé automatique sont variables. De façon générale, ces méthodes construisent un résumé contextuel qui consiste à bâtir un résumé à partir des mots clés qui se situent dans le voisinage immédiat des mots de la requête dans la page. Limiter l'analyse textuelle à cette information débouche donc sur une autre limite : ces quelques lignes sont elles représentatives du contenu de la page web. Le méta moteur de recherche effectue donc une simplification à un double niveau. Il raisonne sur un échantillon de pages web non représentatif et travaille sur un échantillon de texte de chaque page web non représentatif.

4. On a pu observer que l'analyse classificatoire renvoie des grands thèmes non homogènes, qui ne s'inscrivent pas dans les mêmes logiques. Lorsque dans la requête « sciences de l'information et de la communication », on juxtapose @rchiveSIC, SFSIC et journalisme ; on met cote à cote des dimensions qui ne se situent pas au même niveau : publications ouvertes (preprint et postprint) dans le cas d'@rchiveSIC, site institutionnel dans le cas de la SFSIC, thématique dans le cas du journalisme.

3. Approche proposée : validation expérimentale

Nous proposons, dans cette communication, une approche classificatoire qui s'affranchit des principales limites des outils existants. En effet, la méthode que nous avons mise en oeuvre :

-1. considère non pas 1000 pages maximum pour une requête analysée mais l'intégralité des pages disponibles sur le sujet dans le moteur de recherche. Ainsi sur la requête « sciences de l'information et de la communication », nous allons réaliser un traitement non pas sur 1000 pages mais sur 109 000 pages issues de Google.

-2. L'analyse que nous proposons ne se limite pas à traiter l'information contenue dans le titre et le résumé contextuel proposé par le moteur. L'analyse prend en compte l'intégralité du contenu de la page, ainsi que le contenu de pages ayant des formats spécifiques (word, PowerPoint, acrobat).

-3. L'analyse ne renvoie pas une liste de mots non maîtrisée mais propose une liste de termes validés en amont par les experts du domaine.

-4. Tout ceci se paie par un renoncement au principe de travail en temps réel.

Pour arriver à cet objectif, nous mettons en œuvre une chaîne de traitement de l'information qui sera décrite autour d'un exemple pour des raisons pédagogiques. La validation expérimentale que nous allons suivre consiste à s'intéresser à la construction d'une classification automatique d'un corpus constitué de pages web touchant à la problématique des « sciences de l'information et de la communication ». Les outils analysés dans l'état de l'art partent d'un corpus correspondant à une requête d'un internaute, d'un traitement par analyse de contenu pour déboucher sur une classification automatique. Nous allons voir que la méthode que nous proposons est assez différente.

3.1. input de la méthode

Le point de départ de la méthode est une liste de termes caractérisant le domaine. Si on s'intéresse aux « sciences de l'information et de la communication », on peut trouver dans la littérature plusieurs listes de mots clés couvrant le champ des SIC. Nous avons considéré la représentation des SIC que propose @rchiveSIC [3] et qui est rappelée Figure 1. Ainsi notre analyse prend pour point de départ non plus de la requête « sciences de l'information et de la communication » mais une série de requêtes correspondant chacune à un grand domaine d'@rchiveSIC. D'autres listes de départ auraient pu être retenues (mots clés issus du thésaurus Rameau par exemple, champs présents sur le site web de la section 71 du Cnu) ce qui aurait débouché alors sur d'autres classifications.

Figure 1 : liste par grands domaines d'@rchiveSIC

Les résultats finaux de la classification seront constitués des grands domaines d'@rchiveSIC. La méthode que nous proposons consiste à figer en amont une liste de mots clé définis qui constitueront la liste de résultats. Ce qui va nous intéresser c'est non pas trouver des termes à l'issue d'une classification automatique mais réaliser une classification automatique pour positionner ces termes les uns par rapport aux autres. Ainsi l'expert du domaine ne sera pas perdu puisque les documents qu'il visualisera seront encapsulés dans des grands domaines correspondant à une classification humaine réalisée par un expert du domaine.

3.2. la collecte des données :

La limite principale des outils de classification automatique que nous avons étudiés est qu'ils se limitent à l'analyse d'un résumé d'un échantillon de 1000 pages web maximum. Pour autant, n'ayant pas les moyens d'un moteur de recherche, il apparaît difficile de contourner ces limites.

Nous y sommes parvenu par une astuce qui consiste à déporter sur le moteur de recherche une recherche de cooccurrence de termes.

Considérons par exemple deux domaines des SIC : « muséologie » et « Edition électronique ». Si on avait lancé une requête « science de l'information et de la communication », on aurait trouvé ces deux termes associés dans un seul résumé contextuel d'une page renvoyée par le moteur de recherche Google.

CRIC
... pays, de dénommer « **Sciences de l'information et de la communication** », est un **...** de la **muséologie** (code de **...** sur une « Charte d'**édition électronique** ». **...**
www.ir2i.com/colloque_cric.php?url=ethique_deveze - 37k - Résultat complémentaire

Une recherche conjointe de ces deux termes sous Google montre qu'il y a en Janvier 2006 157 pages web contenant ces deux termes.

Nous allons procéder ainsi pour chacun des 21 domaines identifiés par @rchiveSIC. Ces domaines sont étudiés deux par deux à la recherche du nombre de pages web contenant les deux termes simultanément quelque part dans la page. Dans ce travail, il est inutile de récupérer les pages ni les résumés. C'est le moteur de recherche qui réalise le traitement à notre place et qui renvoie, pour chaque association de grands domaines, le nombre de documents ou apparaissent les deux mots clés.

Le travail consiste donc à requêter le moteur de recherche nX(n-1)/2 fois, n correspondant au nombre de grands domaines considérés. Ce travail est concevable manuellement pour un ensemble d'une vingtaine de « grands domaines » mais plus difficilement lorsque la taille de cet ensemble augmente. Pour effectuer ce travail, nous avons conçu un outil automatique qui sollicite le moteur de recherche Google pour chacune des 210 paires de « grands domaines » à la recherche du nombre de réponses ou les deux termes sont co-présents.

Le résultat obtenu se présente sous forme d'une matrice carrée symétrique contenant en ligne et en colonne le nom des 21 grands domaines et, dans chaque cellule, le nombre de pages web associant les grands domaines pris deux à deux.

Un exemple d'une telle matrice est fourni tableau 1. La première diagonale comporte le nombre de pages traitant de chacun des grands domaines pris isolément.

	(Bibliométrie OR scientométrie)	(Cinéma OR art OR esthétique)	"collectivités territoriales"	(communication "information scientifique")	(conflits OR stratégie OR veille)
(Bibliométrie OR scientométrie)	34600	967	184	630	18700
(Cinéma OR art OR esthétique)		793000000	644000	122000	6400000
"collectivités territoriales"			2050000	13500	682000
(communication "information scientifique")				288000	152000
(conflits OR stratégie OR veille)					27000000

Tableau 1 : exemple des cooccurrence de 5 grands domaines

Nous avons calculé tableau 2, pour chaque cellule de cette matrice, le coefficient de Jaccard [7]. Considérons deux des 21 grands domaines identifiés que nous appellerons A et B. La figure 2 illustre, de façon symbolique, le mode de calcul de ce coefficient.

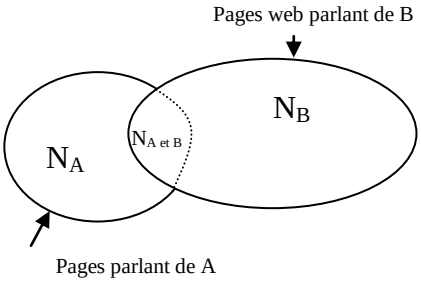

Figure 2 : construction du coefficient de Jaccard

N_A correspond au nombre de pages web parlant du sujet A sans parler du sujet B

N_B correspond au nombre de pages web parlant du sujet B sans parler du sujet A

$N_{A\ et\ B}$ correspond au nombre de pages web parlant à la fois du sujet A et B.

Le coefficient de Jaccard correspond à $N_{A\ et\ B}/(N_{A\ et\ B}+N_A+N_B)$.

Ce coefficient est un indicateur de similitude qui permet de pondérer la valeur de l'association par le nombre de pages web correspondant à chaque terme disjoint.

	(Bibliométrie OR scientométrie)	(Cinéma OR art OR esthétique)	(communication "information scientifique")	(conflits OR stratégie OR veille)
(Bibliométrie OR scientométrie)		0,00%	0,10%	0,03%
(Cinéma OR art OR esthétique)	0,00%		0,01%	0,39%
(communication "information scientifique")	0,10%	0,01%		0,28%
(conflits OR stratégie OR veille)	0,03%	0,39%	0,28%	

Tableau 2 : indicateur de Jaccard

Ainsi le 0.39 % signifie que 0.39% des pages web parlant de (conflit OR stratégie OR veille) ou de (cinéma OR art OR esthétique) parlent de (conflit OR stratégie OR veille) ET de (cinéma OR art OR esthétique)

Cette matrice permet de construire un réseau [10] qui permet de visualiser les interactions les plus fortes entre les grands domaines. Nous avons choisi de conserver toutes les associations pour lesquelles la valeur de Jaccard est supérieure à 2%. Cela représente 46 associations sur les 210 que comporte la matrice soit 21.9 % des interactions. Toutefois, ce sont les interactions les plus fortes. Elles représentent au total 73,9% de la somme des valeurs de la matrice de Jaccard. Le résultat se présente sous la forme d'un réseau représenté figure 2.

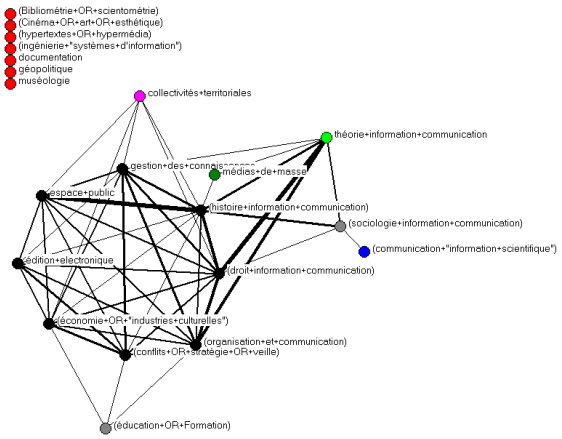

Figure 2 : réseau des interactions

Le réseau se présente sous forme d'une structure réticulaire centrale fortement dense et de 7 sommets isolés. Les sommets de couleur noire correspondent au noyau de taille maximale. L'épaisseur du lien entre deux sommets est proportionnelle au coefficient de Jaccard.

Il est utile de commenter les raisons de l'isolement de ces 7 sommets. Le choix de la distance de Jaccard et du filtre qui a été introduit doit être mis en perspective avec la grande dispersion dans les réponses à ces 21 requêtes comme l'illustre le tableau 2.

depart	nbre de réponse sous Google
(Cinéma OR art OR esthétique)	1574000000
documentation	1482000000
(éducation OR Formation)	324000000
(conflits OR stratégie OR veille)	54000000
(économie OR "industries culturelles")	45400000
(organisation et communication)	33400000
(droit information communication)	23200000
espace public	26600000
gestion des connaissances	19660000
(histoire information communication)	9580000
édition electronique	11000000
théorie information communication	3280000
collectivités territoriales	4720000
médias de masse	3780000
géopolitique	4220000
(sociologie information communication)	2400000
(hypertextes OR hypermédia)	1760000
(communication "information scientifique")	568000
(ingénierie "systèmes d'information")	770000
muséologie	720000
(Bibliométrie OR scientométrie)	69200

Tableau 2 : réponse aux requêtes sous Google (janvier 2006)

Parmi les 21 grands domaines d'@rchiveSIC, certains renvoient sur le web énormément de réponses intégrant du bruit si on se situe dans une perspective de recherche en Info Com. Ainsi la requête (cinéma ou art ou esthétique) comporte-t-elle plus de milliard de réponses. A l'opposé, d'autres requête en présentent beaucoup moins comme bibliométrie ou scientométrie (69200). Cette grande dispersion fait qu'il est parfois mathématiquement impossible que le coefficient de Jaccard soit supérieur à 2%. Il suffit que le nombre de réponses à une des deux requêtes soit 50 fois plus élevé que l'autre pour que le coefficient soit

inférieur à 2%. Pour ces raisons nous proposons d'utiliser un filtre reposant non plus sur le coefficient de Jaccard mais sur une variante du coefficient d'inclusion qui en reprenant les symboliques de la figure 2 se calculerait ainsi :
$N_{A\ et\ B}/(MIN(N_{A\ et\ B}+N_A\ ;\ N_{A\ et\ B}+N_B))$
On a gardé les valeurs de la matrice supérieures à 14% ce qui correspond à 76 interactions réalisant à elles seules 69,4% de la sommes des valeurs de la matrice donc près de 70% de l'information complète du réseau.

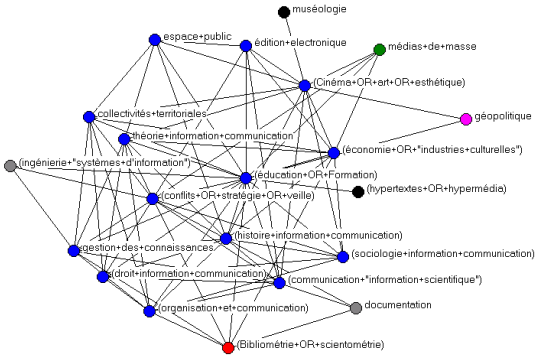

Figure 3 : réseau des interactions avec coefficient d'inclusion

Dans le réseau de la figure trois, les sommets isolés disparaissent. On remarque une densité forte du réseau autour notamment d'un noyau de taille maximale composée des sommets bleus. Le choix de la métrique apparaît comme déterminant dans l'interprétation qui peut être faite des résultats.

Conclusion :
La méthode que nous proposons n'a pas pour vocation à être implémentée dans un méta moteur de recherche temps réel. Plusieurs pistes d'utilisation semblent néanmoins se dessiner. Nous allons en préciser trois de la plus étroite à la plus large
- Tout d'abord, l'approche que nous proposons est une autre manière de fouiller le web. Celui qui recherche de l'information n'est plus contraint par une équation logique limitée à quelques mots clés. Il réalise une recherche en rentrant un ensemble de mots clés libres ou thésaurés. La logique sous jacente est donc celle d'une logique non booléenne. L'approche pourrait s'avérer intéressante dans une problématique de veille sectorielle ou d'intelligence économique
- L'approche peut être utilisée également pour valider des listes de mots clés, thésaurus ou taxinomie sur des corpus web. C'est un peu ce que nous avons fait dans cet exemple en analysant la façon dont la catégorisation a priori d'@rchiveSIC se reflétait sur le web. Sur @rchiveSIC, les déposants ont la possibilité de choisir 3 domaines pour qualifier leurs dépôts. Il serait intéressant d'établir une classification à partir des résultats saisis par les déposants et de comparer les deux résultats.
D'autres analyses de ce type pourraient être envisagées. Une classification automatique réalisée suivant le même principe à partir des descripteurs Rameau des thèses publiés en « sciences de l'information et de la communication » pourrait être envisagée. Un travail analogue conduit à partir des champs de la discipline présentés sur le site web de la section 71 du Cnu pourrait également être effectué.
- Enfin plus largement, cette approche peut se révéler féconde dans des logiques exploratoires de type découverte de connaissances. Si au lieu d'être centré sur un problème, on explore des univers disjoints, il est possible d'identifier des passerelles qui peuvent éveiller la sagacité du chercheur dans une logique de sérendipité [2].

Bibliographie :

[1] Benyu Z.; Hua L.; Yi L.; Lei J.; Wensi X.; Weiguo F.; Zheng C.; Wei-Ying M., (2005), Improving Web Search Results Using Affinity Graph, conference WWW 2005

[2] Ertzscheid O., Gallezot G., "Chercher faux et trouver juste : sérendipité et recherche d'information." 1ère conférence internationale francophone en Sciences de l'Information et de la Communication. Bucarest. Juillet 2003 http://archivesic.ccsd.cnrs.fr/sic_00000689.html

[3]Gallezot G, Rossi C., Chartron G., Noyer J.M. Conception d'une Archive ouverte en SIC : le sens de la technique, Hypertextes hypermedias : créer du sens à l'ère du numérique, H2PTM'03. 24 septembre 2003.

[4]Hartigan J., A. Clustering Algorithms, John Willey, New York, 1975

[5] Huberman B.-A., The Laws of the Web, patterns in the ecology of information, The MIT Press, Cambridge, Massachussets, London, England, 2003.

[6]Prime-Claverie C., Beigbeder M. et Lafouge T. Clusterisation du Web en vue d'extraction de corpus homogènes Communication, INFORSID 2002, 20e congrès informatique des organisations et des systèmes d'information et de décision, Nantes, France, 4-7 Juin:229-242. 06 juin 2002.

[7] Prime C., Bassecoulard E., Zitt M., Co-citations and co-sitations: a cautionary view on an analogy , Proceedings of the 8th International Conference on Scientometrics and Infometrics, ISSI 2001, Sydney, Australia, July 16-20, 2001, p. 529-540.

[8] Rostaing H., Veille technologique et bibliométrie: concepts, outils, applications, Thèse : Université Aix-Marseille III, 1993

[9] Rousseau R. Sitations : an exploratory study , Cybermetrics, 1, (1) 1997

[10] Wasserman S., Faust K. , Social Network Analysis: Methods and Applications. Cambridge, England, and New York: Cambridge University Press, 1994.

[11] White Paper, vivisimo, Tagging versus Clustering for Enterprise Search Engines, http://vivisimo.com/docs/tagging.pdf

Annexe 1 : résultat de la requête « sciences de l'information et de la communication » sous 4 méta-moteurs proposant une classification automatique

EXALEAD	

Annexe 2 : liste non exhaustive de quelques outils de classification automatiques analysés dans l'état de l'art

www.clusty.com/
www.vivisimo.com
www.webclust.com/
turbo10.com/
www.accumo.com/
www.iboogie.com/
www.exalead.com/